Data Curation Network:

Developing and Scaling Research Data Management

Lisa Johnston University of Minnesota
Jake Carlson University of Michigan
Cynthia Hudson--Vitale Washington Univ
Heidi Imker University of Illinois
Wendy Kozlowski Cornell University
Robert Olendorf Penn State University
Claire Stewart University of Minnesota
Mara Blake Johns Hopkins University
Elizabeth Hull Dryad Data Repository
Joel Herndon Duke University
Timothy M. McGeary Duke University

NISO Virtual Conference: Open Data Projects 06-13-2018

Data Curation Network datacurationnetwork.org

Timeline of Reasons to Scale Up RDM and Curation

2011 - National Science Foundation requires Data Management Plans

2013 - OSTP memo requiring Federal agencies with more than $100M in R&D expenditures to develop plans to make the results of federally-funded research freely available to the public—generally within one year of publication.

2016 - NSF and Dept of Energy begin requiring deposit of publications and require data to be made available at expense of research institutions

Why invest in Research Data Management & Curation?

Researchers are faced with a growing number of requirements (and incentives) to ethically share their research data.

Well curated data are more valuable.

The skills and expertise required to curate data cannot be fully automated nor reasonably be provided by a few experts siloed at single institutions.

Why Duke chose to invest in RDM and Curation

2015 - NSF rejects Duke research proposal due to insufficient DMP, specifically due to lack of plans to deposit data

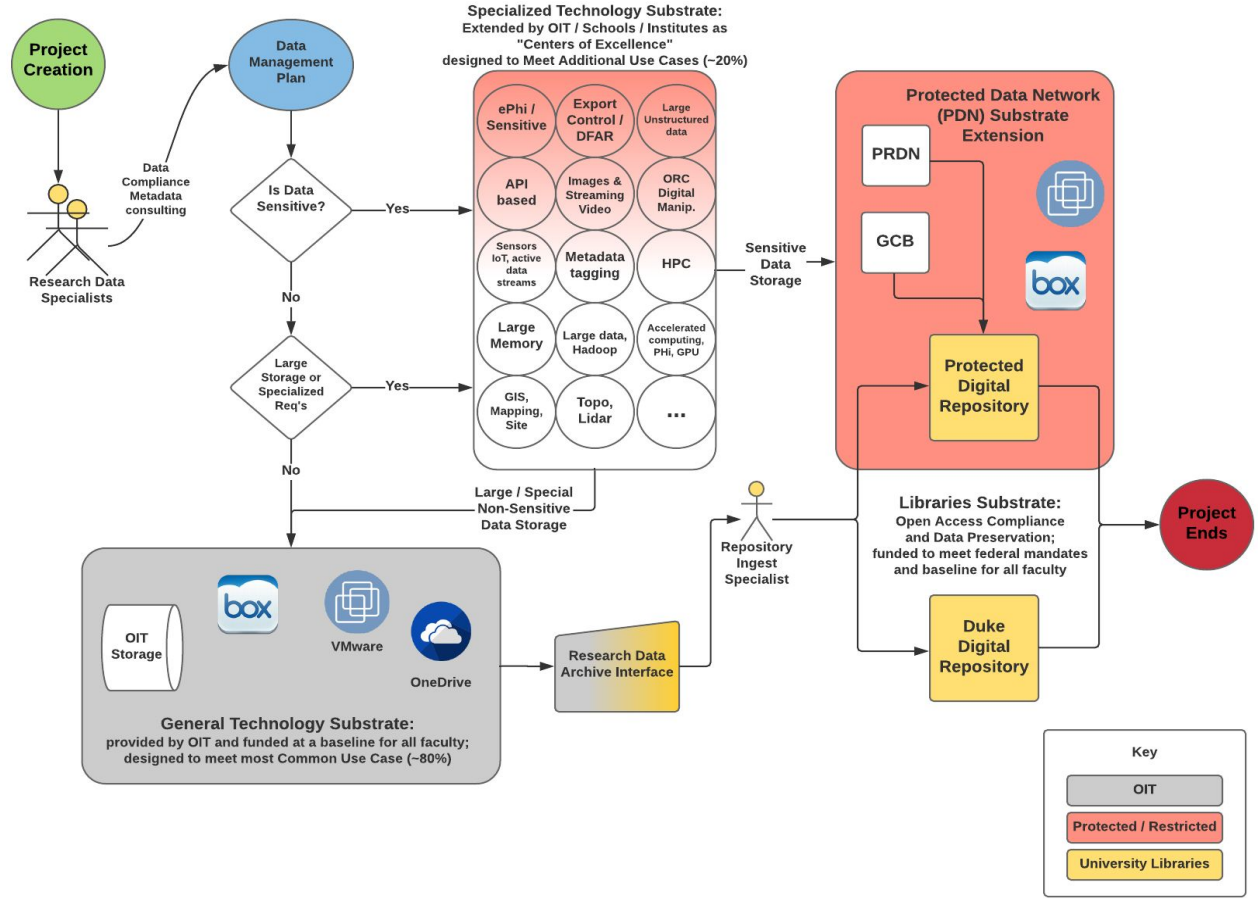
2015 - University Libraries, Trinity College of Arts & Sciences, and Office of Information Technology (OIT) each request significant funding increase for research data storage from Provost

2015 - Provost seeks white paper and charge for Faculty Working Group for Digital Research Data Services

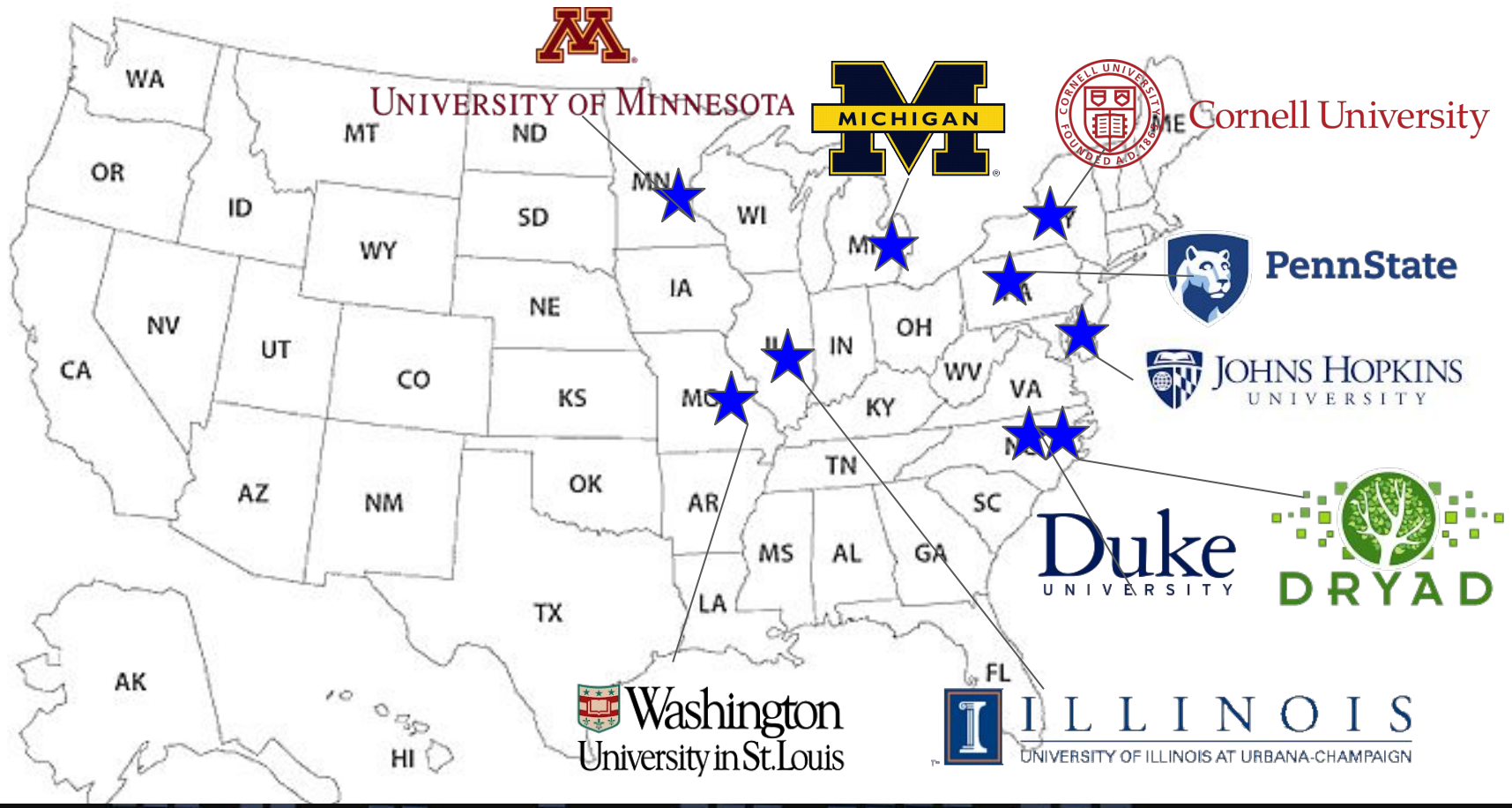
2016 - Interdisciplinary faculty working group meets for 9 months and submits recommendation to Provost

DUKE DIGITAL RESEARCH DATA SERVICE MAP

Tracy Futhey, Tim McGeary | January 22, 2016



The Data Curation Network (DCN) addresses these challenges by collaboratively sharing data curation staff across a network of partner institutions and data repositories.



Data Curation Network datacurationnetwork.org

Steps in building the Network

1. **DCN planning phase research (2016-2017)**

2. **Implementation launch in Spring 2018**
   a. DCN staffing model + Advisory panel
   b. DCN training/networking events
   c. DCN workflow and C-U-R-A-T-E steps
   d. Assessment Plan

3. **Grow the DCN beyond our grant-funded phase to a sustainable entity**
   a. Criteria for new partners
   b. Proposed financial model (alliance curation-as-service)

Planning Phase (2016-2017)

Alfred P. Sloan FOUNDATION

1. **Compared local policy,** technologies, and workflows across the 6 planning phase institutions;
2. **Held six focus groups** with researchers on what data curation activities were important;
3. **Ran controlled pilots** of data curation workflows with 17 data curators to ID issues;
4. **Surveyed the 124** ARL institutions to gauge support for data curation services;
5. **Researched cost recovery** models for sustainable data curation and repository services;
6. **Held information exchanges** with leaders of successful collaboration projects;
7. **Analyzed one-year** of data types, disciplines, frequency, and curation levels needed vs taken).

Planning Phase (2016-2017)

Alfred P. Sloan FOUNDATION

How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

Baseline Assessment

	Pre-ingest Curation?		Mediated vs Self-deposit?		Post-ingest curation			
Workflow Steps by Institution	Consult only	Staging Area for deposit	Mediated deposit	Self-deposit	As needed	Review metadata only	Review files and metadata	Add DOI
Minnesota	X			X			X	X
Cornell	X		X*	X			X	X*
Illinois	X			X			X*	X
Michigan	X			X			X*	X*
Penn State	X			X				
Wash U	X		X	X			X	X

"Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions' Data Repository and Curation Services." Journal of eScience Librarianship 6(1): e1102. https://doi.org/10.7191/jeslib.2017.1102.

Planning Phase (2016-2017)

Alfred P. Sloan FOUNDATION

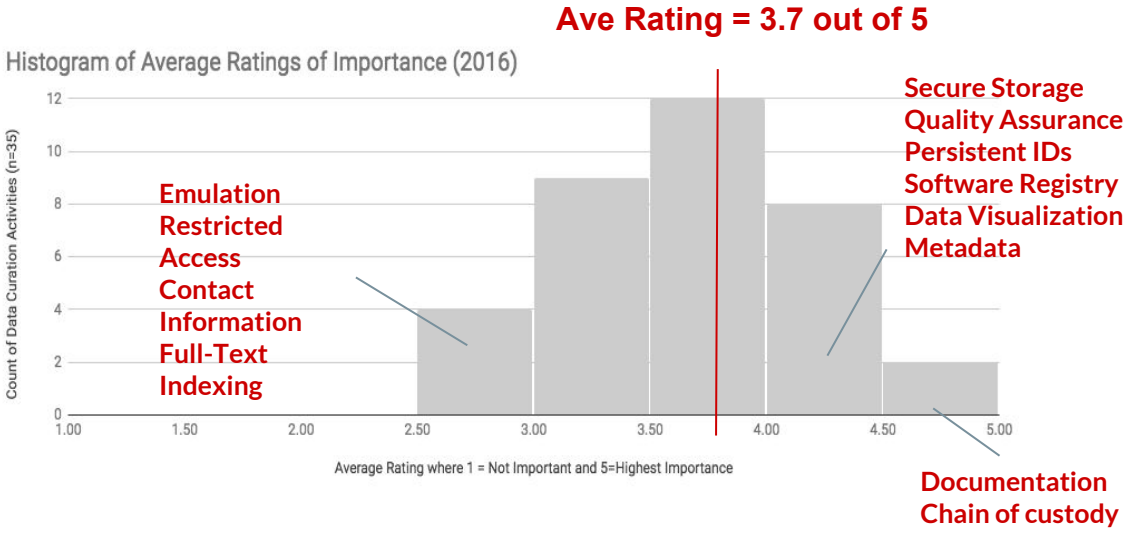
How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

Researcher Focus Groups (n=91)



(in press) "How Important Are Data Curation Activities to Researchers? Gaps and Opportunities for Academic Libraries," *Journal of Librarianship and Scholarly Communication.*

Planning Phase (2016-2017)

Alfred P. Sloan FOUNDATION

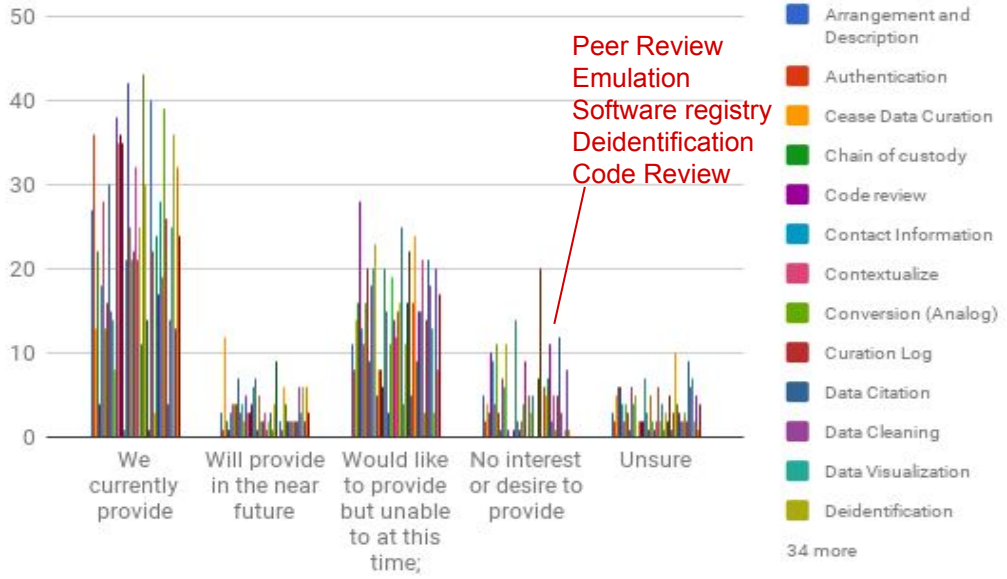
How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

ARL Institutions Survey (n=80)



SPEC Kit #354: Data Curation. Association of Research Libraries (ARL). May 2017.
http://publications.arl.org/Data-Curation-SPEC-Kit-354/~~FreeAttachments/Data-Curation-SPEC-Kit-354.pdf

Planning Phase (2016-2017)

Alfred P. Sloan FOUNDATION

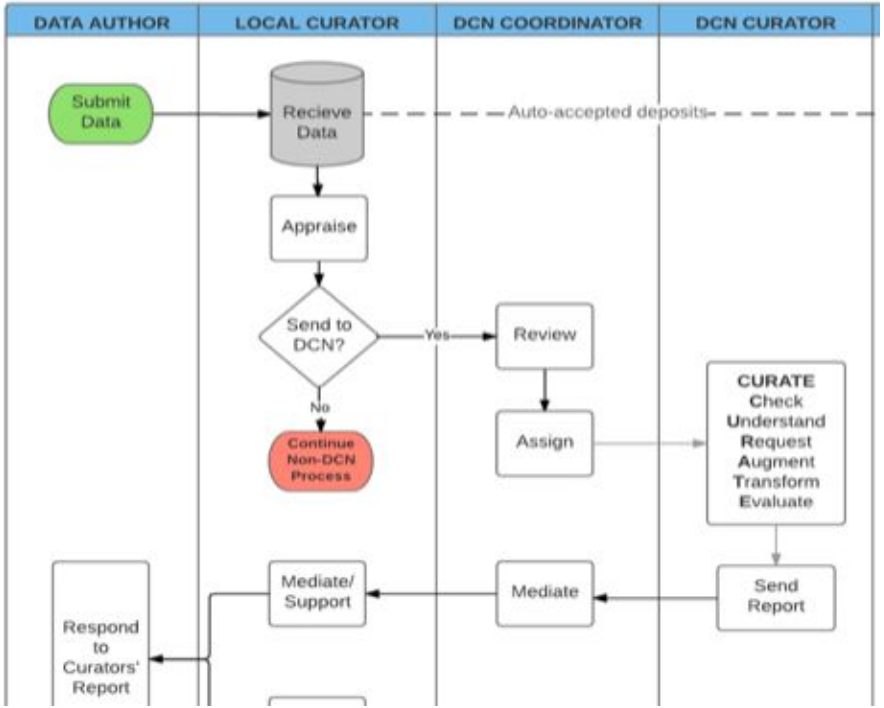
How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

Data Curator Pilots



Planning Phase (2016-2017)

Alfred P. Sloan FOUNDATION

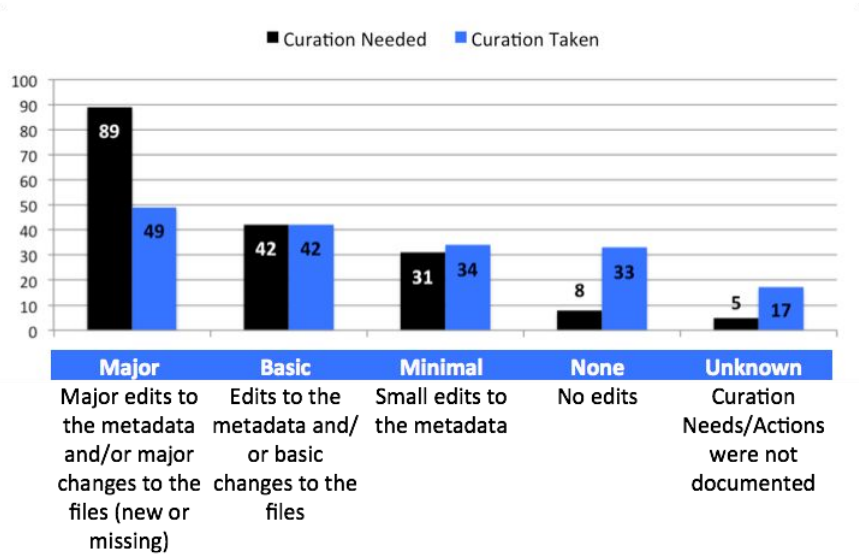
How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

One Year Metrics Tracking (n=175)



Report: "Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data" (2017), http://hdl.handle.net/11299/188654.

Planning Phase (2016-2017)

Alfred P. Sloan FOUNDATION

What are the measure of success?

How can we grow and sustain the Network beyond the grant-funding period?

Information Exchanges

2CUL

DURASPACE

portage
SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA

SEAD

University of California CDL California Digital Library

TDL.ORG TEXAS DIGITAL LIBRARY

ISB International Society for Biocuration

CURATING for REPRODUCIBILITY

FDA U.S Food and Drug Administration

DCN Implementation

9 Institutions
- 8 Academic Libraries
- 1 General Data Repository

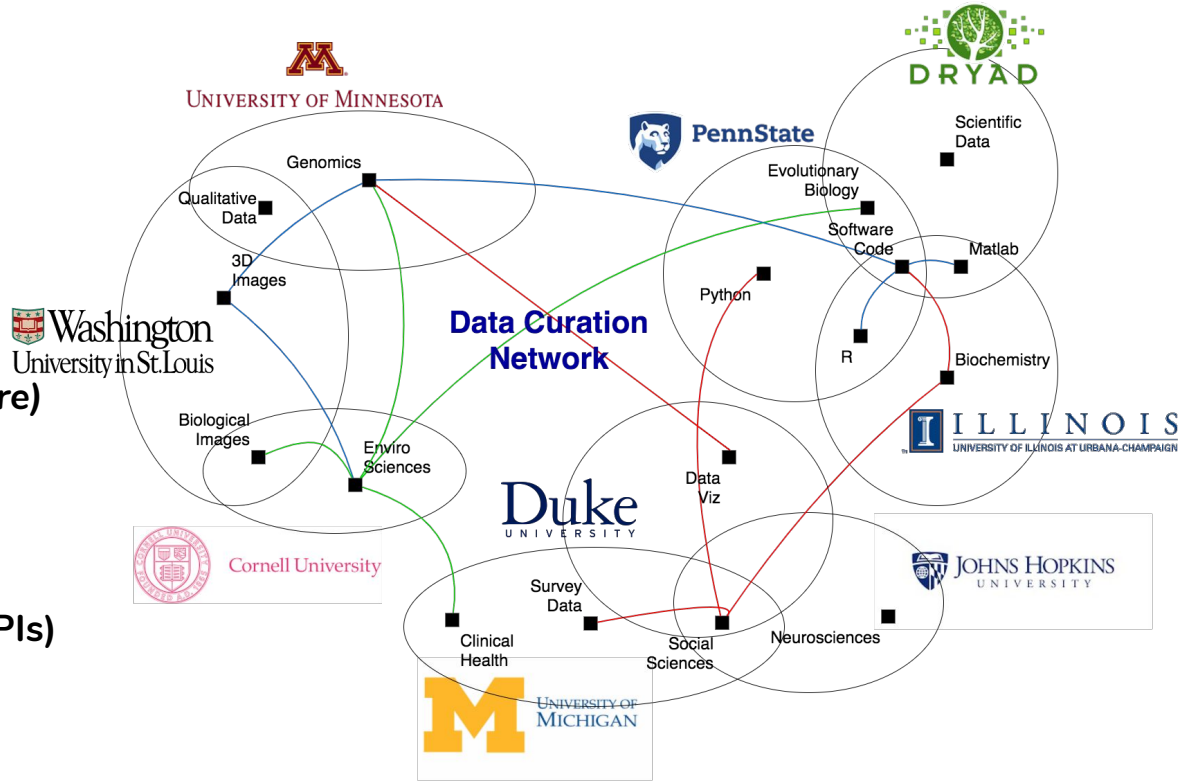
19 Data Curators

1 Project Coordinator ***(new hire)***

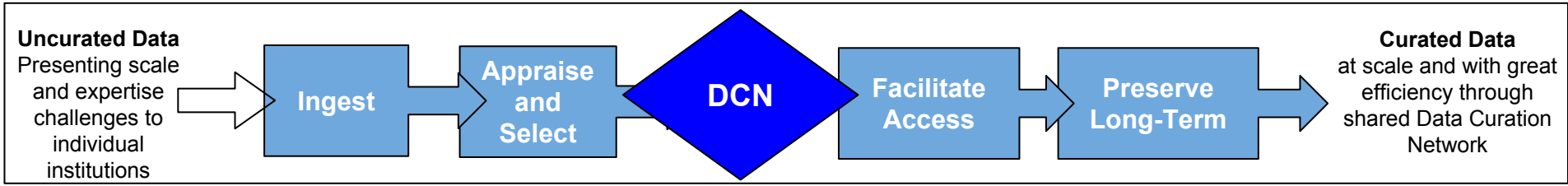
1 Program Director **(PI)**

8 DCN Representatives **(CO-PIs)**

2 Admin Leads

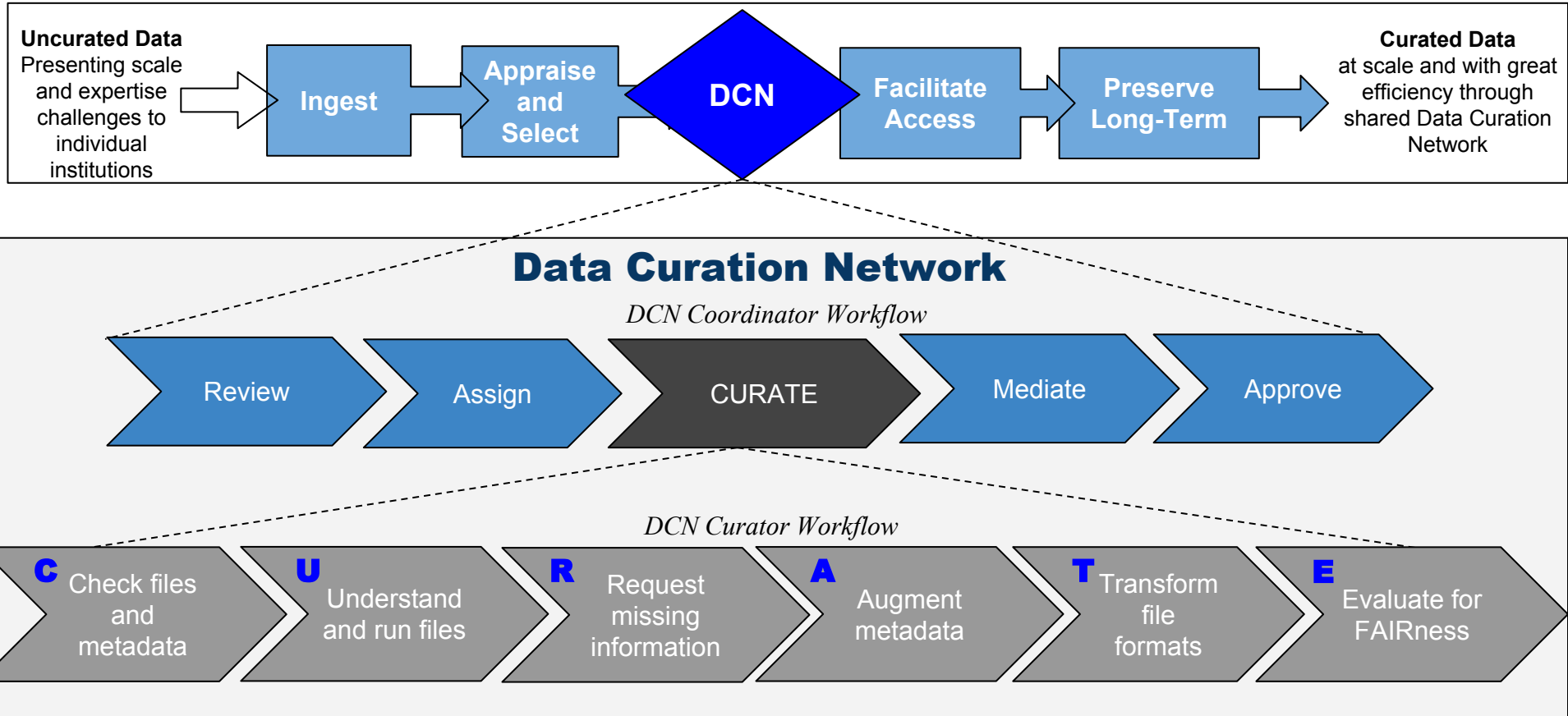


DCN Workflow



- Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)
- Local institution maintain full responsibility for all technical functionality (eg. storage) and authority for local decision-making (what to ingest, how long to retain, etc.)
- Seamlessly integrates into all repository systems (Samvera, Fedora, DSpace, etc.)

DCN Workflow



CURATE Steps in DCN Workflow

DCN Curators will take **CURATE** steps for each data set, that includes:

- **C** **Check** data files and read documentation
- **U** **Understand** the data (try to), if not…
- **R** **Request** missing information or changes
- **A** **Augment** the submission with metadata for findability
- **T** **Transform** file formats for reuse and long-term preservation
- **E** **Evaluate** and rate the overall submission for FAIRness.

DCN Implementation (2018-2020)

Assessment Plan (two-prong)

Is a networked approach to curating research data more efficient?

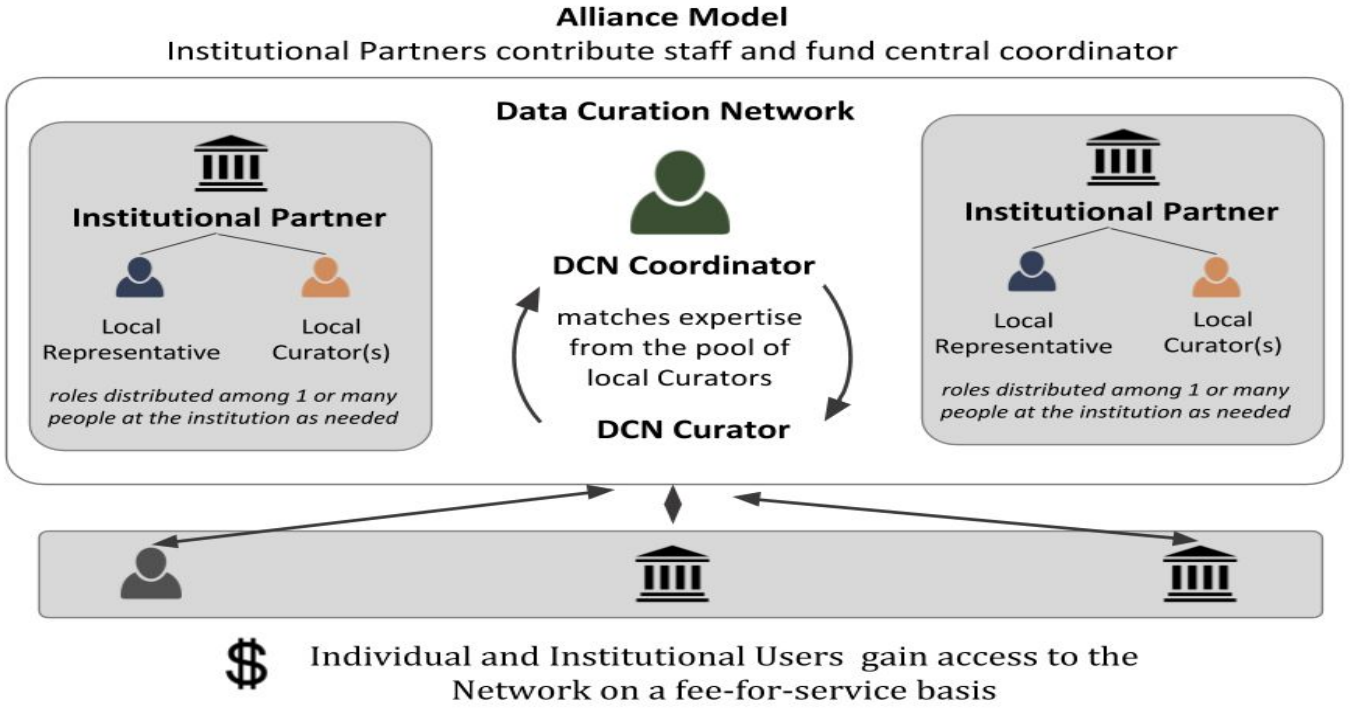
- Number of datasets
- Frequency (high-volume time periods, etc.)
- Variety (data file formats; range of disciplines)
- Efficiency (time, costs)

Are curated data are more valuable?

- Track reuse indicators (download counts, citations, alt-metrics)
- Implement a DCN registry
- Apply badges and metadata to signal that data sets curated by the DCN are FAIR.

*In Year 3, the DCN will begin transitioning to a **self-sustaining service model** where institutional and disciplinary partners contribute data curation staff and central operations costs are offset by users of the Network.*

Data Curation-as-service



Value proposition

Stakeholder	*Benefits*
Academic libraries with existing data curation services	Gain access to data curation expertise in more disciplines/formats than locally available
Academic libraries with limited to no resources for data curation services	Are able to provide critical new data curation services when local resources are limited (without needing to hire);
Disciplinary- and general-subject data repositories	Receive better, more valuable data submissions from DCN partner institutions and customers; Have potential to partner with the DCN to expand the scope of curation support for new and/or less frequently encountered data types

Data Curation Network FAQ

- Do researchers actually value these services?
- Won't researchers curate their own data?
- Is it another community of practice?
- Aren't you all large research libraries? How about other kinds of libraries?
- Can't some of this (data curation) be automated?
- Why aren't you also sharing a repository?
- Why charge $ for this?

What other questions do you have?

Data Curation Network: Thank you

Timothy M. McGeary
Duke University
Associate University Librarian for
Digital Strategies & Technology
tim.mcgeary@duke.edu

NISO Virtual Conference: Open Data Projects 06-13-2018

Data Curation Network datacurationnetwork.org